Semantic Network Analysis

van Atteveldt, W.H.

2008

document version
Publisher's PDF, also known as Version of record

Samenvatting (Dutch Summary)

In onze democratie spelen de media een belangrijke rol als podium en recensent voor het politiek schouwspel. De meeste informatie over politieke actoren en thema's komen de kiezers via de media te weten. Het is dus van groot belang om te begrijpen hoe de media functioneren en welke invloed zij — bedoeld of onbedoeld — uitoefenen op de gedachten en handelingen van het publiek.

In de sociale wetenschappen wordt het onderzoeken van de inhoud van communicatieboodschappen, zoals televisie-uitzendingen of krantenartikelen, *inhoudsanalyse* genoemd. In de kwantitatieve stroming bestaat deze inhoudsanalyse doorgaans uit het bepalen van de frequentie van interessante variabelen, bijvoorbeeld hoe vaak de verschillende politici worden genoemd, of hoe vaak de berichtgeving op een bepaalde manier *geframed* wordt.

Alhoewel met deze handmatige en 'thematische' inhoudsanalyse succesvol onderzoek is gedaan, kleven er ook een aantal nadelen aan. Handmatige inhoudsanalyse is arbeidsintensief, en daardoor duur en tijdrovend. Een ander nadeel is dat de verzamelde gegevens vaak zeer nauw aansluiten op de onderzoeksvraag. Dit maakt het moeilijk om gegevens opnieuw te gebruiken: als bepaalde thema's of frames op een bepaalde manier geteld zijn, is het niet mogelijk deze tellingen te gebruiken voor een andere of zelfs een enigszins gewijzigde onderzoeksvraag.

Zoals beschreven in hoofdstuk 2 is een alternatieve methode voor inhoudsanalyse de semantische netwerkanalyse. In deze methode wordt niet volstaan met het tellen van bepaalde thema's, maar wordt elke boodschap ontleed in een semantisch netwerk van relaties tussen de relevante actoren en onderwerpen: wie steunt wie, welke standpunten worden in-

genomen, gaat het goed met de actoren en hoe ontwikkellen de issues zich? Aan de hand van deze semantische netwerken wordt vervolgens de oospronkelijke onderzoeksvraag beantwoord, bijvoorbeeld hoe vaak het nieuws op een bepaalde manier *geframed* wordt.

Door deze loskoppeling van tekstanalyse en onderzoeksvraag, met het semantisch netwerk als 'halffabrikaat', is het mogelijk om verschillende onderzoeksvragen te beantwoorden met dezelfde gegevens, of gegevens die in verschillende onderzoeken zijn verzameld te combineren tot grotere databestanden. Dit biedt een efficientievoordeel en maakt het makkelijker om grote databestanden te ontwikkelen en verschillende theorieën direct te vergelijken op basis van deze bestanden.

Ook semantische netwerkanalyse heeft echter nadelen. Ten eerste is het handmatig extraheren van de netwerken uit teksten een nog grotere inspanning dan handmatige thematische analyse. Daarnaast is het in de praktijk vaak lastig om bestaande netwerken te koppelen, omdat de actoren en onderwerpen in die netwerken vaak niet overeenkomen: nieuwe issues komen op, politici veranderen van rol en partij, nieuwe partijen worden gesticht en ontbonden. Tenslotte is het analyseren van semantische netwerken vaak een complexe bezigheid, wat het moeilijker maakt om deze methode in te zetten.

De drie inhoudelijke delen van dit proefschrift beschrijven een aantal methoden en technieken die zijn ontwikkeld om deze nadelen van semantische netwerkanalyse te verminderen. Deze technieken zijn ruwweg onder te verdelen in extractie (deel II), representatie en bevraging (deel III), en systeembeschrijving (deel IV).

Voor de *extractie* van semantische netwerken zijn een aantal technieken uit de computerlinguïstiek gebruikt om het automatisch extraheren van netwerken te vergemakkelijken. Hoofdstuk 5 kijkt naar hoe het samen voorkomen (co-occurentie) van actoren en onderwerpen kan worden gebruikt als een eerste 'associatief' semantisch netwerk. Dit wordt geïllustreerd met een toegepast onderzoek naar de berichtgeving over Islam en terrorisme. Dit toont aan dat op een relatief simpele manier associatieve netwerken geëxtraheerd kunnen worden uit grote verzamelingen teksten, en dat die netwerken interessante inzichten kunnen opleveren over die teksten.

In hoofdstuk 6 wordt dit verrijkt door de grammaticale analyse van zinnen te gebruiken om een onderscheid te maken tussen bron, handelende actor, en lijdend voorwerp. Een *case study* laat zien dat deze informatie gebruikt kan worden om de media-autoriteit van actoren te onderzoeken. De betrouwbaarheid van deze methode wordt gemeten door op zowel meet- als analyseniveau een vergelijking te maken tussen de automatische analyse en een eerdere handmatige analyse. Hieruit blijkt dat, alhoewel de techniek zeker nog verbeterd kan worden, de betrouw-

baarheid op analyseniveau hoog genoeg is voor onmiddelijk gebruik in de sociale wetenschap.

Ten slotte wordt in hoofdstuk 7 door gebruik van *machine learning* op basis van bestaande gegevens de lading (positief – negatief) van de relatie bepaald. Deze methode blijkt de handmatige analyses redelijk te benaderen en werkt significant beter dan een simpelere 'baseline' methode. Ook in een aantal concrete *case studies* heeft de methode een goede correlatie met de uitkomsten van een handmatige analyse, en in de meeste gevallen is de correlatie hoog genoeg om op de resultaten te kunnen vertrouwen voor automatische inhoudsanalyse.

Naast de extractie kijkt dit proefschrift naar het *representeren en bevragen* van de geëxtraheerde semantische netwerken. Hoofdstuk 8 beschrijft hoe de taal RDF van het Semantische Web gebruikt kan worden om zowel het semantisch netwerk zelf te representeren als de benodigde achtergrondkennis om dit netwerk te analyseren. In die achtergrondkennis staat bijvoorbeeld van politici wanneer ze lid waren van welke partij en welke functies zij vervulden. Van de onderwerpen is opgenomen tot welke hoofdonderwerpen ze behoren.

Hoofdstuk 9 beschrijft een relatief simpele 'querytaal' waarmee onderzoekers of geïnteresseerden het semantisch netwerk kunnen bevragen om bepaalde patronen te zoeken. De resultaten hiervan kunnen getoond en gevisualiseerd worden op zowel geaggregeerd niveau als op het niveau van de oorspronkelijke artikelen waarin het patroon voorkwam. De geaggregeerde gegevens kunnen weer gebruikt worden voor verdere kwantitatieve analyse.

Het laatste deel van dit proefschrift, hoofdstuk 10, bevat de *systeembeschrijving* van de infrastructuur die is ontwikkeld als deel van dit onderzoek. *AmCAT*, de Amsterdam Content Analysis Toolkit, is een systeem om documenten zoals krantenartikelen op te slaan, te doorzoeken, en te prepareren voor analyse. *iNet* is een programma voor handmatige semantische netwerkanalyse dat gekoppeld is aan het AmCAT systeem om het beheer van grootschalige codeeroperaties te vergemakkelijken en om te zorgen dat de coderingen gekoppeld zijn aan de documenten en ontologie in AmCAT.

De verschillende delen van dit proefschrift beslaan een breed scala aan onderwerpen: extractie van semantische netwerken met natuurlijke-taalverwerking, representatie en analyse van deze netwerken met technieken uit de kennisrepresentatie, en een concreet systeem voor grootschalige handmatige en automatische inhoudsanalyse. Samen genomen betekent dit een belangrijke stap voorwaarts voor semantische netwerkanalyse, wat het makkelijker maakt om deze techniek in te zetten in de communicatiewetenschap.